Applied and Environmental Microbiology, Nov. 2009, p. 7229–7242 Vol. 75, No. 22
0099-2240/09/$12.00 doi:10.1128/AEM.00857-09


# Identification of *Bacillus anthracis* by Using Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry and Artificial Neural Networks[▽]

Peter Lasch,[1]* Wolfgang Beyer,[2] Herbert Nattermann,[3] Maren Stämmler,[1] Enrico Siegbrecht,[1] Roland Grunow,[3] and Dieter Naumann[1]

*P25[1] and ZBS2,[3] Robert Koch-Institut, Nordufer 20, D-13353 Berlin, Germany, and Institute for Environmental and Animal Hygiene and Veterinary Medicine at the University of Hohenheim, Garbenstraße 30, D-70599 Stuttgart, Germany[2]*

Received 15 April 2009/Accepted 14 September 2009

**This report demonstrates the applicability of a combination of matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry (MS) and chemometrics for rapid and reliable identification of vegetative cells of the causative agent of anthrax, *Bacillus anthracis*. *Bacillus* cultures were prepared under standardized conditions and inactivated according to a recently developed MS-compatible inactivation protocol for highly pathogenic microorganisms. MALDI-TOF MS was then employed to collect spectra from the microbial samples and to build up a database of bacterial reference spectra. This database comprised mass peak profiles of 374 strains from *Bacillus* and related genera, among them 102 strains of *B. anthracis* and 121 strains of *B. cereus*. The information contained in the database was investigated by means of visual inspection of gel view representations, univariate *t* tests for biomarker identification, unsupervised hierarchical clustering, and artificial neural networks (ANNs). Analysis of gel views and independent *t* tests suggested *B. anthracis*- and *B. cereus* group-specific signals. For example, mass spectra of *B. anthracis* exhibited discriminating biomarkers at 4,606, 5,413, and 6,679 Da. A systematic search in proteomic databases allowed tentative assignment of some of the biomarkers to ribosomal protein or small acid-soluble proteins. Multivariate pattern analysis by unsupervised hierarchical cluster analysis further revealed a subproteome-based taxonomy of the genus *Bacillus*. Superior classification accuracy was achieved when supervised ANNs were employed. For the identification of *B. anthracis*, independent validation of optimized ANN models yielded a diagnostic sensitivity of 100% and a specificity of 100%.**

Members of the genus *Bacillus* are rod-shaped bacteria that exhibit catalase activity and can be characterized as endospore-forming obligate or facultative aerobes. The genus *Bacillus* contains two important groups of bacteria named after *B. subtilis* and *B. cereus*. The best-characterized member of the former group is *B. subtilis*, a renowned model organism for genetic research. Other group members, like *B. pumilis*, *B. licheniformis*, *B. atrophaeus*, and *B. amyloliquefaciens*, exhibit a high degree of phenotypic similarity and are thus not easily distinguishable (15).

The *B. cereus* group comprises a number of closely related bacteria, some of which interfere with human health. Bacteria classified as *B. cereus* are occasionally associated with food poisoning (16, 28), while *B. thuringiensis* is primarily an insect pathogen because of its ability to produce toxins that have been widely used for the biocontrol of insect pests (28, 30). A third member of the *B. cereus* group, *B. anthracis*, is the causative agent of anthrax and is highly relevant to human and animal health. Other members of the *B. cereus* group are *B. mycoides*, *B. pseudomycoides*, and *B. weihenstephanensis* (4, 15).

*B. anthracis* is a possible agent in biological warfare and bioterrorism. Its applicability as a biological warfare agent was made apparent by an accidental release from a Soviet military facility in Sverdlovsk (1, 10). Also, the well-publicized mailing of *B. anthracis* spores in the United States, which caused 18 confirmed cases of cutaneous and inhalational anthrax and an additional 4 suspected cases of cutaneous anthrax (3, 22), demonstrated that *B. anthracis* may become a threat from terrorist groups (10).

Rapid detection of *B. anthracis* may be challenging because of its great genetic similarity to other species of the *B. cereus* group (10) and the difficulties of phenotypic differentiation of *B. cereus* group members (15). There is some controversy in the literature regarding the taxonomy of the *B. cereus* group. Indeed, some authors state that *B. anthracis*, *B. cereus*, and *B. thuringiensis* are one species with various virulence plasmids for the toxin pXO1 and the capsule pXO2 of *B. anthracis* and the insecticidal toxin of *B. thuringiensis* (10, 19). Other authors do not support this opinion and suggest the presence of even more species within the group (21).

Matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) intact-cell mass spectrometry (ICMS) has been suggested as a rapid, objective, and reliable technique for bacterial identification (8, 13, 23, 25, 38). As a proteomic technique, ICMS of whole bacterial cells, or cell lysates, relies on the reproducible detection of microbial protein patterns and thus delivers information complementary to genotypic or phenotypic test methods. With the pattern-matching approach, microbial identification is achieved by comparing experimental

* Corresponding author. Mailing address: P25, Robert Koch-Institut, Nordufer 20, D-13353 Berlin, Germany. Phone: 49-30-45472405. Fax: 49-30-45472606. E-mail: LaschP@rki.de.
▽ Published ahead of print on 18 September 2009.

mass spectra with a collection of mass spectra of known organisms. This requires the compilation of large databases of bacterial reference spectra but has the advantage that an extensive knowledge of biomarker identities is not required. Another advantage of the pattern-matching approach is that genus- and species-specific procedures or consumables are not required, i.e., the same methodology can in principle be applied to all kinds of microorganisms (multiplex advantage).

It is thus believed that ICMS offers the possibility to systematically investigate the diversity of bacterial subproteomes, complementing existing methodologies of bacterial characterization. This potential and the need for a rapid, objective, and reliable microbial identification technique that does not rely on nucleic acid detection and the availability of an MS-compatible inactivation protocol for highly pathogenic biosafety level 3 microorganisms and bacterial endospores (26) prompted us to systematically study the MALDI-TOF MS profiles of *Bacillus* strains and to establish a database of bacterial mass spectra. In the present work, we describe strategies of spectral analysis that allow the identification and validation of group- and species-specific sets of biomarkers. Using unsupervised hierarchical cluster analysis (UHCA) and supervised artificial neural network (ANN) analysis, we also demonstrate how microbial spectra can be employed to establish an MS-based methodology for rapid, objective, and reliable identification of the target species, *B. anthracis*.

## MATERIALS AND METHODS

**Microbial strains and isolates.** Most of the bacterial strains originated from the strain collections of the Institute for Environmental and Animal Hygiene and Veterinary Medicine at the University of Hohenheim, Stuttgart, Germany, and the Robert Koch-Institut, Berlin, Germany. Isolates of *B. anthracis* from these culture collections were confirmed as target sequences of both virulence plasmids (where applicable) by culture diagnosis on blood agar, gamma phage susceptibility, and the presence of a chromosomal signature sequence by real-time PCR, as proposed previously (5, 40). Further *Bacillus* strains were obtained from the German Collection of Microorganisms and Cell Cultures GmbH (Braunschweig, Germany), from the strain collection of the Microbial Ecology Group at the Technical University Munich (M. Ehling-Schulz), and from the Bundeswehr Research Institute for Protective Technologies and NBC Protection (WIS) in Munster (B. Niederwöhrmeier). An overview of the *Bacillus* strains and isolates used throughout the study is provided in Table 1.

**Sample preparation.** Starting from a pure single colony on blood agar, *Bacillus* cultures were prepared by growing each strain for two passages under aerobic conditions on LB agar (Merck) for 24 h at 37°C. Cells were harvested by transferring the equivalent of three full blue plastic loops (equivalent to about 30 μl) from each agar plate into 20 μl of sterile water. The bacterial material was resuspended by vortexing it. Sample preparation and sample inactivation were carried out by using the modified trifluoroacetic acid (TFA) inactivation protocol for highly pathogenic microorganisms (26). Briefly, 80 μl of pure TFA (Uvasol; Merck) was added to 20 μl of the bacterial water suspensions. After being gently shaken for 30 min, the solutions were diluted 10-fold with high-performance liquid chromatography grade water (Mallinckrodt Baker B.V., Deventer, The Netherlands). No centrifugation or filtration was carried out. The sample dilutions were checked for sterility. For MALDI-TOF MS, 2 μl of the microbial dilution was then mixed with 2 μl of a 12-mg/ml α-cyano-4-hydroxycinnamic acid (HCCA) solution (Bruker Daltonics, Bremen, Germany). Preparation of the HCCA solution was achieved by dissolving HCCA in $TA_2$, a 2:1 (vol/vol) mixture of 100% acetonitrile and 0.3% TFA. One microliter of the sample-HCCA mixture was spotted onto ground-steel sample targets from Bruker Daltonics.

**MALDI-TOF MS.** Mass spectra of microbial extracts were collected using an Autoflex mass spectrometer from Bruker Daltonics. The instrument was equipped with a slightly defocused $N_2$ laser operating at 337 nm at pulse rates up to 20 Hz. The pulse ion extraction time was 200 ns. Measurements were carried out in the linear mode using an acceleration voltage of 20.00 (ion source 1) or 18.45 (ion source 2) kV. The lens voltage was 6.70 kV. The mass spectra were stored in the low and intermediate mass range between 0.5 and 2 kDa and between 2 and 20 kDa, respectively. Insulin (5,734.52 Da), ubiquitin 1 (8,565.89 Da), cytochrome *c* (12,360.97 Da), and myoglobin (16,952.31 Da) were used as external calibrants, enabling a mass accuracy of about 300 ppm. At least 600 individual laser shots were coadded for each spectrum.

**Data analysis.** Mass spectra were acquired by employing Bruker's Flex Control software package (v. 2.4; Bruker Daltonics). The analysis of mass spectral patterns was carried out using Matlab (Mathworks Inc., Natick, MA)-based software developed in house. The first steps of data analysis comprised spectral preprocessing procedures, such as smoothing, baseline correction, and intensity normalization. For advanced visualization of microbial biomarkers, the Matlab routines allowed the generation of simulated gel views from preprocessed mass spectra. These gel views displayed peak intensities gray-scaled with abscissa values as mass/charge ratios (*m*/*z*) and spectral numbers as the ordinates.

The Matlab routines allowed the preparation of peak lists from preprocessed mass spectra. These peak lists were automatically generated in predefined mass ranges using custom-designed concepts. For example, the procedure of peak list generation makes use of a sigmoidal weighting function that was introduced to correct for the decreased sensitivity of MALDI-TOF MS in the high mass range. Furthermore, the peak detection algorithm produces for each mass spectrum a peak-ranking table. Thus, peak lists contained not only *m*/*z* and intensity information, but also a parameter of the relative importance of individual mass signals. Peak ranks were found to be useful for multivariate spectral analyses, either as alternatives for peak intensities or for data reduction purposes. For the latter case, either the number of peaks included in subsequent analyses or thresholds of peak importance could be specifically defined. The peak lists were further converted to so-called bar code spectra. In bar code spectra, only information about peak presence, or absence, is employed, while the peak amplitude information (intensity) is omitted. The bar code spectra served as inputs for UHCA and analysis by ANNs.

Cluster analysis was carried out by means of distance or *D* values obtained from normalized Pearson's product-momentum correlation coefficients (20) with Ward's algorithm as the clustering method (36). *D* values are regarded as interspectral distance measures that vary between 0 and 2,000 (0 to 1,000 for positively correlated data). Within the context of the present work, UHCA of mass spectra from *Bacillus* isolates was carried out on the basis of the 30 most relevant mass signals taken from the mass range of 2 to 12 kDa (see above).

The in-house-developed Matlab routines also contained an interface with the NeuroDeveloper software package from Synthon (Synthon Analytics GmbH, Heidelberg, Germany). The NeuroDeveloper combines modules for feature selection, ANN model development (including modular ANN models), and ANN-based classification. A detailed description of these software functions can be found elsewhere (35). In the present study, the general strategy of ANN analysis included teaching and optimization of a neural network, followed by testing the classifier with independent (external) test spectra. Teaching and internal validation were done with mass spectra with known class assignments, which have been combined in so-called teaching and internal-validation subsets. The accuracy of the classification was subsequently determined by challenging the classifier with external test spectra that were kept totally separate during model development. The teaching subset of this study contained spectra from 296 bacterial samples. From this number, the NeuroDeveloper software allowed the automatic selection of 20% for internal validation. The number of randomly selected spectra of the external test set was 127. For network teaching, the bar code mass spectra of the microbial database were generated using peak tables with 75 mass signals per spectrum and a point spacing of 700 ppm. Subsequently, the spectra were imported into the NeuroDeveloper program, and 100 spectral features were chosen by the built-in UNIVAR feature selection method, which is based on univariate F statistics. The ANN consisted of a three-layer feed-forward multilayer perceptron (MLP) ANN with connected layers of 100 input, 7 hidden, and 3 output neurons with shortcut connections. Teaching of the ANNs was carried out by utilizing the resilient back-propagation (rprop) function (29).

Matlab-based software was also employed to calculate average mass spectra and for univariate statistical tests (two-sample *t* tests of independent samples).

## RESULTS

Figure 1 displays representative MALDI-TOF mass spectra from four different *Bacillus* species: *B. anthracis* (A), *B. cereus* (B), *B. thuringiensis* (C), and *B. licheniformis* (D). The spectra were acquired from microbial samples prepared by using the modified TFA sample inactivation/protein extraction proce-

TABLE 1. Overview of the bacterial strains used in this study

| Species | Strain (description)[a] |
|---|---|
| *B. anthracis* | DSM 2071, ATCC 11966, ATCC 14185, ATCC 14578, CDC 1014, Vollum, 15 Stamatin vaccine, STI-1 vaccine, STI-1 vaccine black, STI-1 vaccine blue, forest anthrax strain,[b] forest anthrax strain [A],[b] forest anthrax strain [B][b]; A3–A12, A16, A18, A19, A22–A25, A27–A29, **A30**, A32–A39, **A40–A42**, A43–A45, **A46**, A47, A49, A59–A65, A66, A67, *A68*, A69, A73, A80–A85, **A86**, A87, A88 black, A88 blue, A89 black, A89 blue, *A91*, **A92**, A93, A100–A103, A106–A108, A110–A113, A120–A122, A127–A130, **A131–A133**, A135, A136 |
| *B. amyloliquefaciens* | DSM 7 |
| *B. atrophaeus* | DSM 675, DSM 2277, ATCC 9372 |
| *B. badius* | CCEB 623, B106, B107 |
| *B. cereus* | DSM 31, DSM ID 92-340, DSM ID 92-477, DSM 318, DSM 336, DSM 345, DSM 351, DSM 428, DSM 487, DSM 508, DSM 609, DSM 612, DSM 626, DSM 679, DSM 689, DSM 1274, DSM 2299, DSM 2301, DSM 2302, DSM 3101, DSM 4218, DSM 4312, DSM 4490, DSM 6941, DSM 8438, ATCC 6464, ATCC 7004, ATCC 9139, ATCC 9634, ATCC 10876, ATCC 10876a, ATCC 10987, ATCC 11778, ATCC 12826, ATCC 13061, ATCC 14579, ATCC 33018, ATCC 33019, NCTC 6474, NCTC 9939, NCTC 10320, NCTC 10989, NCTC 11143, CCM 88, CCM 99, CCM 2543, WSBC 10028, WSBC 10030, WSBC 10286, WSBC 10288, WSBC 10441, WIS 29, WIS 311, BO 365, BO 366, BO 368, BO 372, BO 488, BO 490, BO 493, 301101RA0432, B3, B10, B16, B33, B108–B110, B147, B252–B262, B272–B274, B292, *B278*, *B282*, *B293–B300*, *B302–B306*, *B308–B311*, B329, B331, *B316*, *B338*, Hohenheim; Paciflor 1, 10, 43, 92, and 111, Paciflor[c]; Toyocerin 5, 8, 9, 14, 16, 19, and 23, Toyocerin[c] |
| *B. circulans* | DSM 11, DSM 1315, DSM 7257, ATCC 9966, NCTC 5846, CCEB 626 |
| *B. coagulans* | ATCC 10545, NCTC 3992, CCEB 631, B112, B114, B340 |
| *B. firmus* | DSM 12, DSM 1530, ATCC 8247, ATCC 14575, CCM 37, CCEB 627 |
| *B. lentus* | DSM 9, DSM 5221, ATCC 10840, CCM 35 |
| *B. licheniformis* | DSM 13, DSM 603, DSM 12369, ATCC 10716, ATCC 12759, ATCC 14580, CCM, 864, 301101RA0433, C (mucilaginous), B62, B115, B116–B119 |
| *B. megaterium* | DSM 32, DSM 90, DSM 509, DSM 333, ATCC 4342, *B2*, *B17*, *B120*, *B333* |
| *B. mycoides* | DSM 2048, WSBC 10279, WSBC 10291, WSBC 10360, WSBC 2641, *B103*, *B124*, *B335*; *B. cereus* var. mycoides |
| *B. pseudomycoides* | DSM 1244 |
| *B. pumilus* | DSM 27, DSM 492, DSM 355, DSM 361, DSM 13835, B31, CCM 340, MM 171 |
| *B. sphaericus* | DSM 28, DSM 396, ATCC 4525, ATCC 10208, ATCC 14577, CCM 436, NCTC 2608, NCTC 5896, NCTC 7582, NCTC 7585, NCTC 9602, NCTC 10338 |
| *B. subtilis* | DSM 10, DSM 347, DSM 618, DSM 1092, DSM 1087, ATCC 6633, CCM 878; CCM 110, *B19*, *B61* (all *B. globigii*); MM 508, 301101RA0436; B, E, I (all rough); WIS no. 491, B1, B4, B6, B7, B12, B14, B20, B28, B32, B158; Glaxo B129, B130 |
| *B. thuringiensis* | DSM 350, DSM 2046, WSBC 28001, WS 2614, WS 2621, WS 2623, WS 2734, B8; *B. thuringiensis* subsp. *israelensis* WIS no. 494 |
| *B. weihenstephaniensis* | DSM 11821, WSBC 10204, WSBC 10207, WSBC 10295, WSBC 10363, WSBC 10387 |
| *Brevibacillus brevis* | DSM 5619, *B336* |
| *Brevibacillus laterosporus* | ATCC 9141 |
| *Paenibacillus alvei* | NCTC 3349, CCM 203 |
| *Paenibacillus macerans* | ATCC 8244 |
| *Sporosarcina pasteurii* | ATCC 11859 |
| *Virigibacillus pantothenticus* | ATCC 14576, B104, B105, B125 |
| *Bacillus* spp. | *BW-A*, *BW-B*, **B21**, *B22*, *B301*, *B307*, B313, B314, B318, B319, *B321*, *B324*, *B334* |

[a] Animal isolates are in boldface, environmental isolates are in italics, and human isolates are underlined. DSM, Deutsche Sammlung von Mikroorganismen; ATCC, American Type Culture Collection; CCM, Czech Collection of Microorganisms; CDC, Centers for Disease Control and Prevention; A and B, strain collection at University Hohenheim (W. Beyer); NCTC, National Collection of Type Cultures; WS, Weihenstephan general collection; WSBC, Weihenstephan *Bacillus* collection; WIS, Bundeswehr Research Institute for Protective Technologies and NBC Protection; CCEB, Culture Collection of Entomogenous Bacteria.

[b] From reference 24.

[c] Collection of probiotic *B. cereus* strains at RKI.

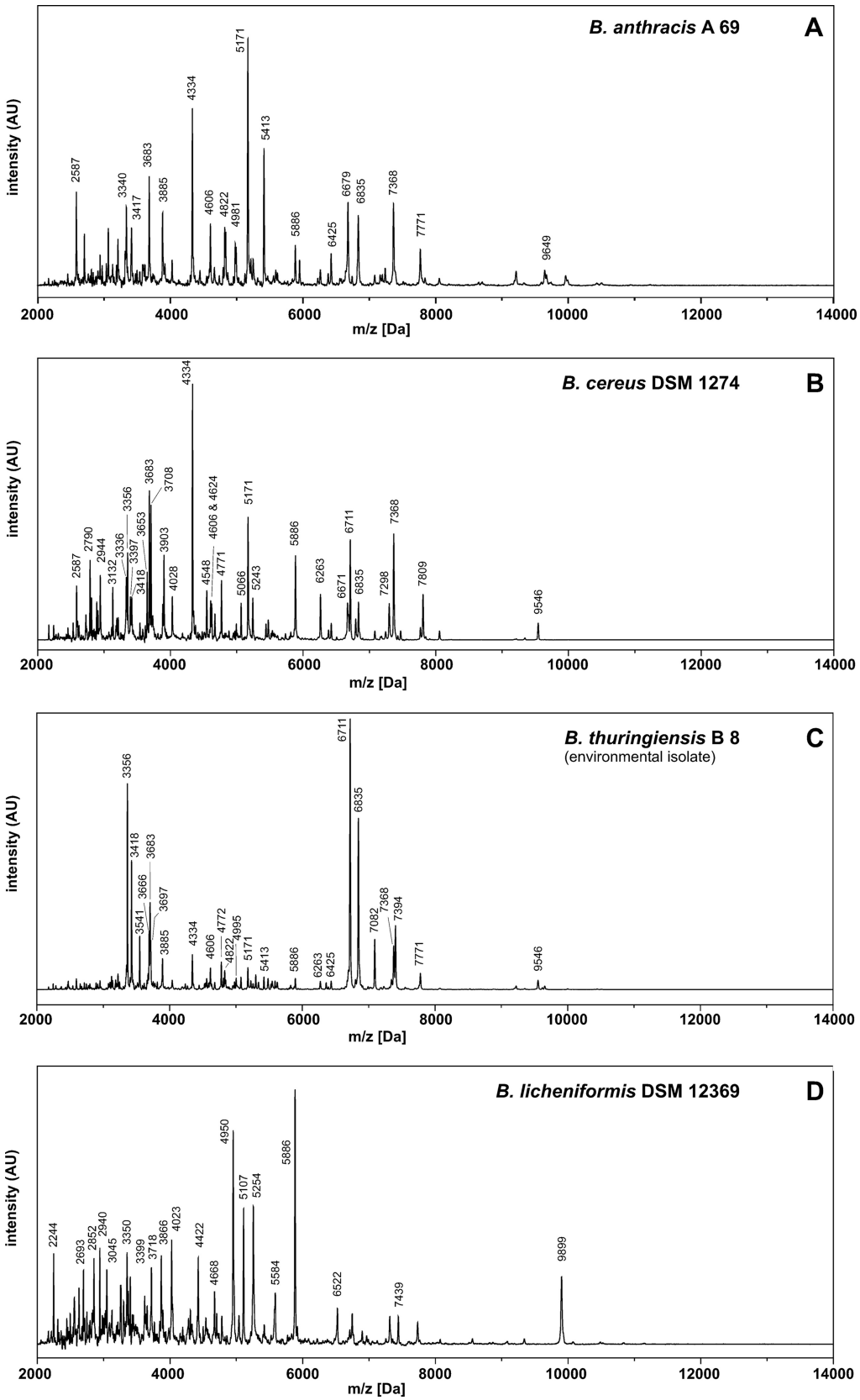


FIG. 1. Representative MALDI-TOF mass spectra of four different *Bacillus* species in the mass range of 2 to 14 kDa. The spectra were acquired from microbial samples using a recently published sample preparation and inactivation protocol for highly pathogenic microorganisms (26) that is based on treatment with 80% TFA. The mass spectra have been baseline corrected, smoothed, and vector normalized. AU, arbitrary units.

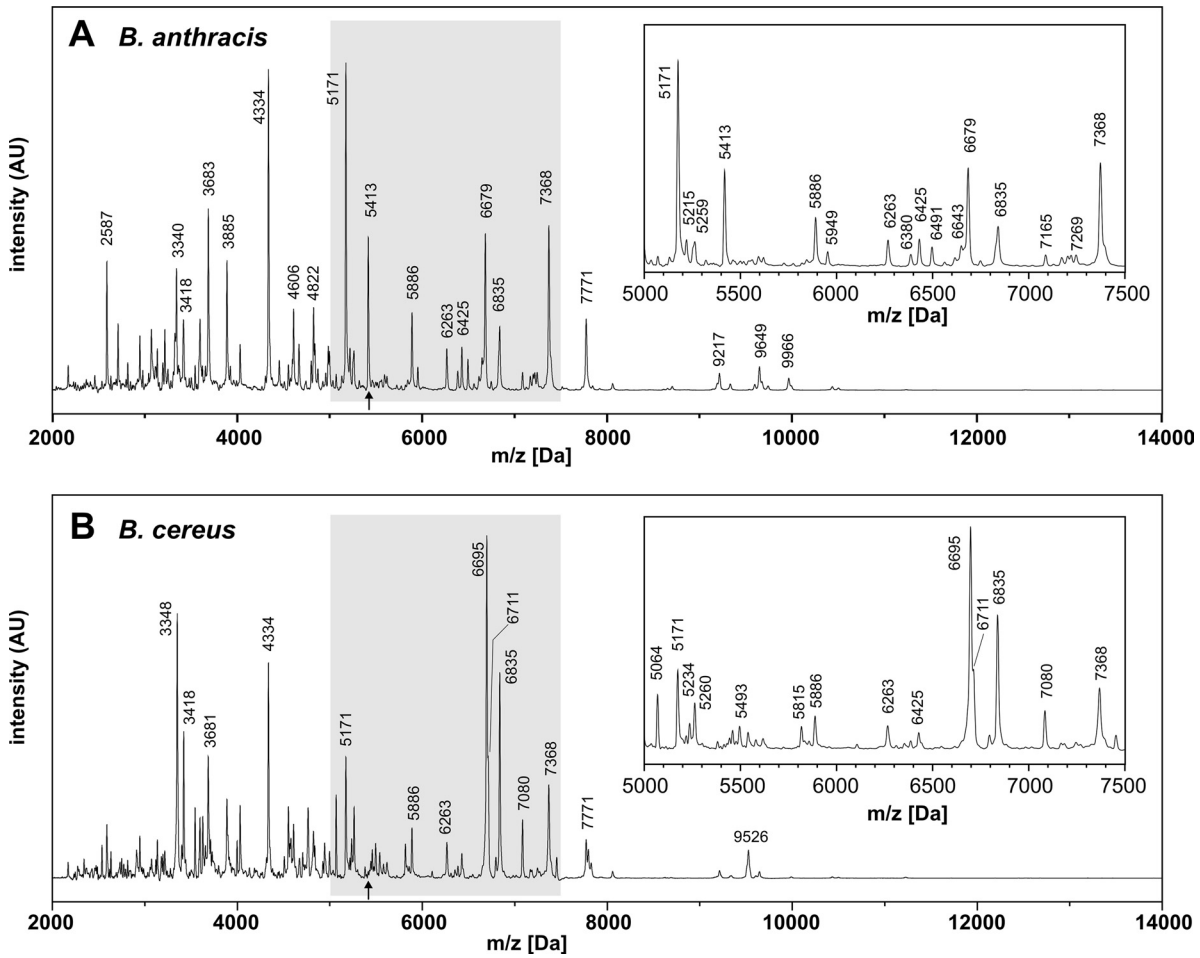


FIG. 2. Mean MALDI-TOF mass spectra of (A) *B. anthracis* (105 spectra) and (B) *B. cereus* (147 spectra). The spectra were obtained by averaging the respective experimental mass spectra from all available strains (see Table 1 and the text for details). AU, arbitrary units.

dure, which included sample treatment with 80% TFA for 30 min (26). The mass spectra shown in Fig. 1 were preprocessed as outlined in Materials and Methods. The spectra demonstrate a relatively high signal-to-noise ratio, which typically permits the detection of 50 to 100 mass peaks per spectrum. A closer inspection of these and other MALDI spectra showed a high degree of distinctness for most of the mass spectral patterns. The example in Fig. 1 also shows the presence of species- or group-specific mass signals. For example, mass spectra from the *B. cereus* group members *B. anthracis*, *B. cereus*, and *B. thuringiensis* (Fig. 1A to C) display a number of group-specific peaks at 3,683, 4,334, 5,171, 5,886, and 7,368 Da. Depending on the degree of sporulation, individual mass patterns (particularly the intensities) may vary considerably. This important aspect is discussed below.

The spectra in Fig. 2 were obtained by averaging preprocessed MALDI-TOF mass spectra from all available *B. anthracis* (105 spectra) and *B. cereus* (147 spectra) strains. At first glance, the two mean spectra appear to be quite different. A more detailed analysis, however, revealed a high number of common mass signals of these closely related species. In fact, there were very few mass peaks that could serve as potential candidates for species differentiation. For example, the spectrum of *B. anthracis* (Fig. 2A) shows a peak at 5,413 Da that has no counterpart in the mean spectrum of *B. cereus* (Fig. 2A and B). For *B. anthracis*, other potential biomarkers can be found at 6,679 Da and for *B. cereus* at 6,695 and 6,711 Da. These markers were previously assigned to species-specific small acid-soluble proteins (SASPs) (6, 11, 17) and are discussed below. At this point, it should be noted that the analysis of mean spectra alone is certainly not sufficient for identifying species-, or group-specific biomarkers. For this purpose, a statistical analysis of the complete mass spectral database is required.

In order to give an overview of the database of microbial MALDI spectra, we prepared so-called gel views. These representations were obtained by converting spectral peak intensities to gray scales, which were then plotted as functions of the *m*/*z* values. It is important to note that the gel view representations were created on the basis of preprocessed spectra, i.e., spectra that had been smoothed, baseline corrected, and vector normalized. The spectra appear as rows, while vertical lines indicate reproducible mass peaks. In Fig. 3A, the gel view shows all 423 spectra obtained within the context of this study in the mass range of 2.5 to 8 kDa. The mass spectra of *B. anthracis* are shown in Fig. 3A at lines 1 to 105 (105 spectra).

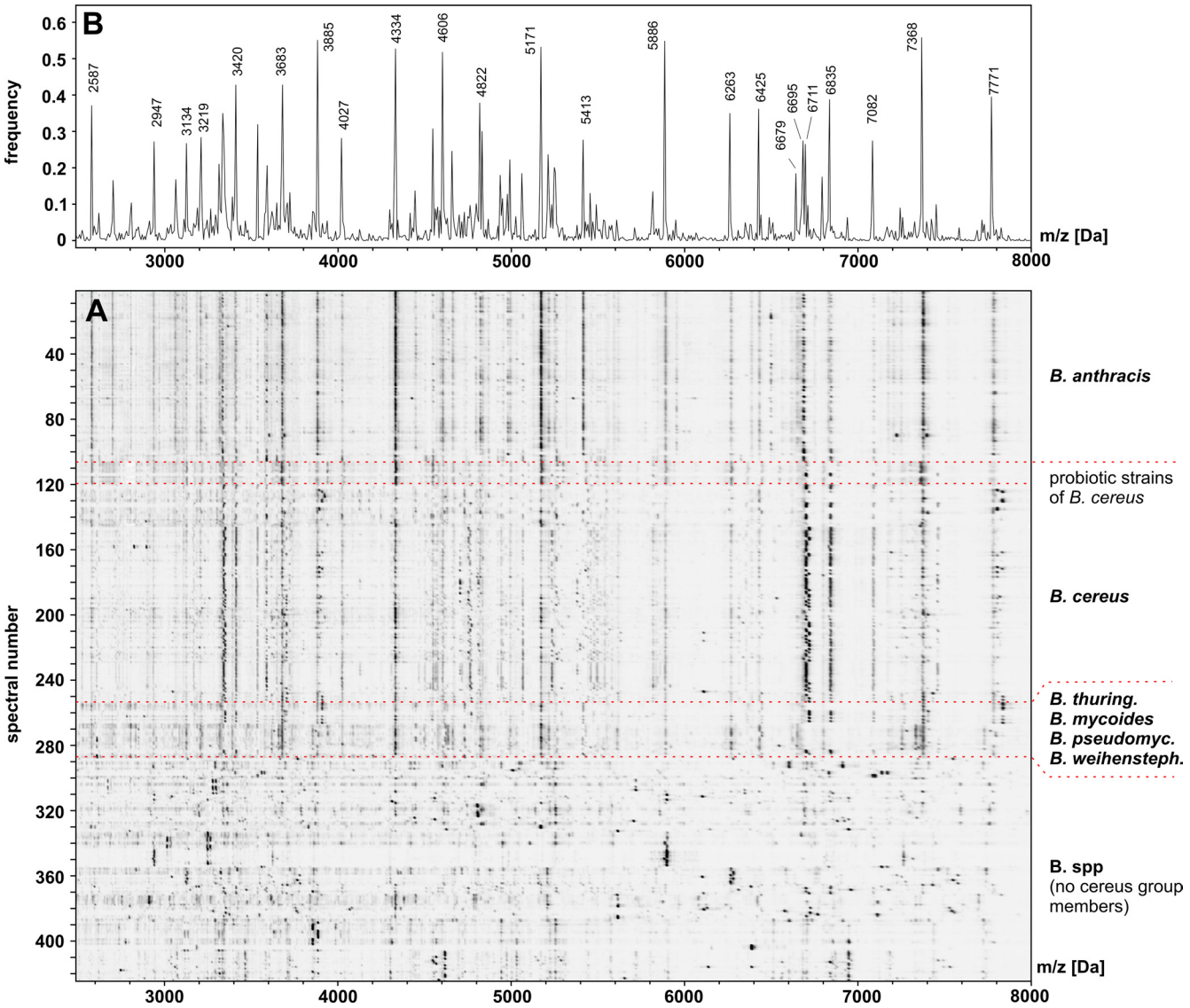


FIG. 3. (A) Gel view representation of mass spectra from *Bacillus* strains in the mass range of 2.5 to 8 kDa. In this gel view, the intensities of 423 MALDI-TOF mass spectra were gray scaled and plotted as functions of the *m/z* values. The spectra have been baseline corrected, smoothed, and vector normalized. (B) Frequency chart of mass peaks obtained from the MALDI-TOF mass spectra shown in panel A. Mass/charge ratio-dependent peak frequencies were obtained by the formula $n^+/n$, with $n^+$ being the number of spectra exhibiting a peak at the current mass/charge ratio. The variable *n* represents the total number of spectra (423). From each spectrum, the 30 most relevant mass peaks were obtained (see the text for details).

The spectra of *B. cereus* strains form lines 106 to 252 (147 spectra). Lines 253 to 286 display spectra from the remaining *B. cereus* group members, *B. thuringiensis*, *B. mycoides*, *B. pseudomycoides*, and *B. weihenstephanensis* (34 spectra). Lines 287 to 423 (137 spectra) were generated from spectra of the non-*B. cereus* group members (18 species) (Table 1). The relative frequencies of the individual mass signals are shown in Fig. 3B. This peak frequency plot was obtained from mass peak tables with 30 entries per individual mass spectrum and illustrates the dominance of mass signals from *B. cereus* group members (see below).

Analysis of the gel view in Fig. 3A demonstrated a high degree of spectral reproducibility and indicated a relatively high degree of similarity between spectra from *B. cereus* group members and, conversely, large-scale heterogeneity for spectra from the remaining species. The analysis of the gel view also suggested the presence of species- and group-specific biomarkers. For example, MALDI-TOF mass spectra of *B. cereus* group members consistently exhibited signals at 3,683, 4,334, 5,171, 5,886, and 7,368 Da, among others. These signals were only rarely detected in mass spectra of the remaining species and therefore can be considered *B. cereus* group biomarker candidates. An example of a species-specific signal is the mass peak at 5,413 Da in spectra of *B. anthracis*. This signal was found to be present in most of the mass spectra (92%) of *B. anthracis* and was already mentioned when discussing the mean spectrum of *B. anthracis* (Fig. 2A). As the gel view shows, the peak at 5,413 Da is, with a few exceptions, absent in the spectra

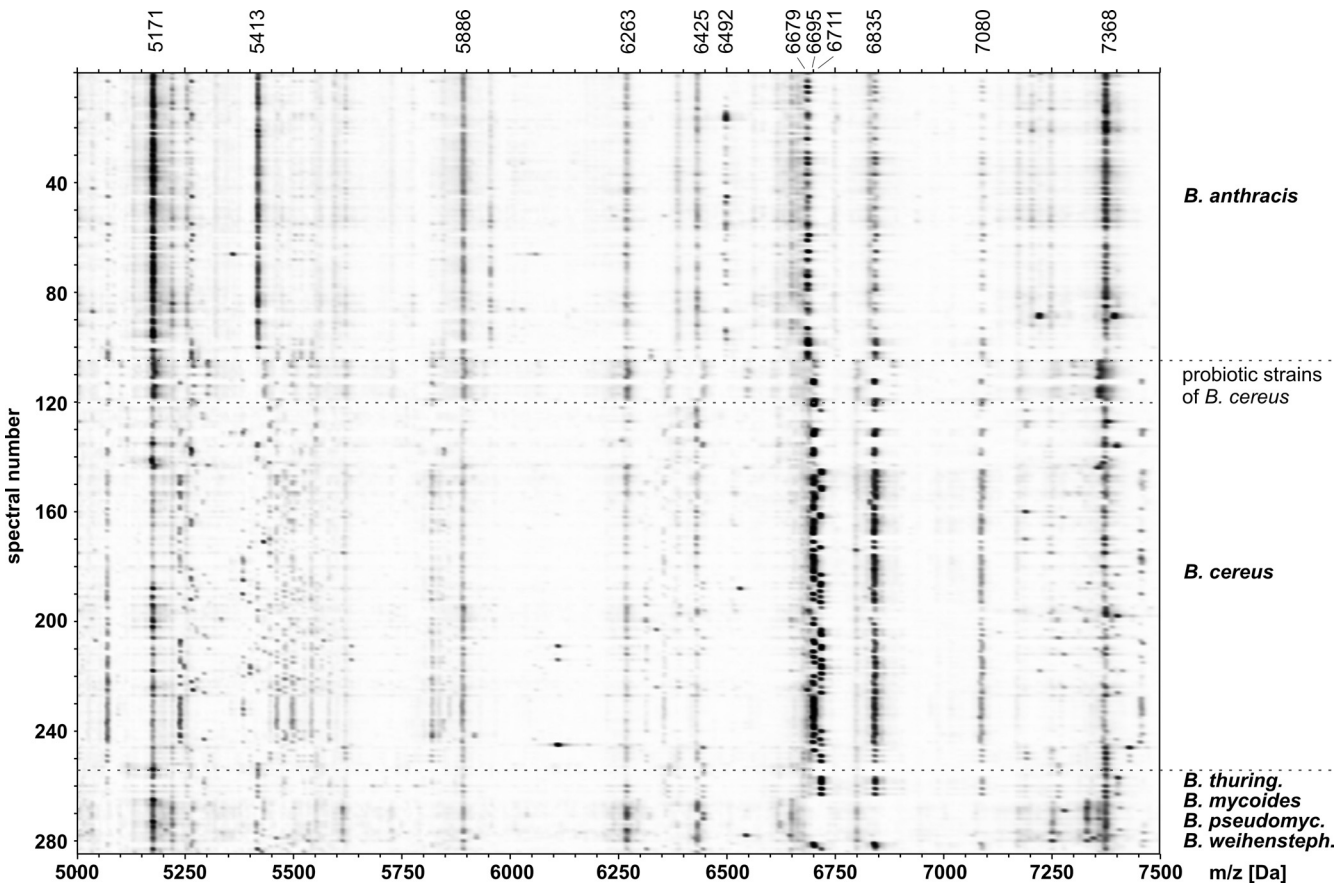


FIG. 4. Gel view of mass spectra from the *B. cereus* group in the diagnostically relevant mass range of 5 to 7.5 kDa. Mass spectra from *B. cereus* group members exhibit marker peaks at 5,171, 5,886, and 7,368 Da (arrows). Peaks at 6,679 Da (*B. anthracis*) or 6,695 and 6,711 Da (*B. cereus*) are due to the presence of small amounts of bacterial endospores containing SASPs (see the text for details).

of all other *Bacillus* species. The exceptions are the strains *B. cereus* DSM 4490, DSM 609, ATCC 12826, and B248; *B. thuringiensis* DSM 350, DSM 2046, and B8; and *B. mycoides* DSM 2048 and B335.

In order to illustrate selected aspects of species discrimination and identification by MALDI-TOF MS with particular emphasis on *B. anthracis*, the gel view in Fig. 4 shows the spectra from the *B. cereus* group in the diagnostically important region between 5 and 7.5 kDa. This representation demonstrates again the existence of common *B. cereus* group mass peaks at 5,171, 5,886, and 7,368 Da (Fig. 3A). More importantly, the illustration also suggests the existence of *B. anthracis*-specific peaks at 5,413 and 6,492 Da and a SASP signal at 6,679 Da. The last peak is of particular interest, because only two strains of *B. cereus* (B248 and B292) exhibit signals at this *m*/*z* value. Most of the strains of *B. cereus* exhibit strong mass peaks close to the 6,679-Da peak, either at 6,695 Da or at 6,711 Da. The latter signals also arise from SASPs and can be observed in some MALDI spectra of *B. thuringiensis* and *B. mycoides*.

**UHCA.** UHCA is a data-driven multivariate classification technique that was employed in this study to investigate the relatedness between the microbial mass patterns. As outlined above, UHCA was carried out on the basis of bar code spectra obtained from peak tables with 30 entries per mass spectrum. Figure 5 shows the dendrogram of a cluster analysis carried out with a mass spectral database of 423 spectra complemented by 14 additional spectra from *B. cereus* group members with unclear species assignments. The dendrogram clearly shows the existence of two main clusters, with a large cluster (I) formed by spectra of the *B. cereus* group and a second cluster (II) containing spectra from non-*B. cereus* group species. Cluster I consists of two main clusters, each subdivided into three subclusters (Ia to If). It is interesting that the MALDI spectra of *B. anthracis* form two of these clusters (Ib and Id), with the majority of the *B. anthracis* strains grouped in cluster Id. Spectra of *B. cereus* and the other species of the *B. cereus* group constitute the four remaining subclusters of cluster I. Cluster Ic contains exclusively spectra of *B. cereus*, while clusters Ia and Ie also comprise spectra of *B. thuringiensis* (Ia) or *B. mycoides*, *B. pseudomycoides*, *B. thuringiensis*, and *B. weihenstephanensis* (Ie). A special situation is observed for mass spectra of probiotic strains of *B. cereus*, which are all located in a separate cluster (If) clearly separated from the other *B. cereus* strains.

Cluster II seems to be less structured, an observation that is backed by the heterogeneity of mass peaks in the gel view in Fig. 3A. It should be mentioned, however, that strains of the *B. subtilis* group members *B. firmus*/*B. lentus* (IIb), *B. licheniformis* (IIc), and *B. subtilis*/*B. atrophaeus*/*B. amyloliquefaciens*

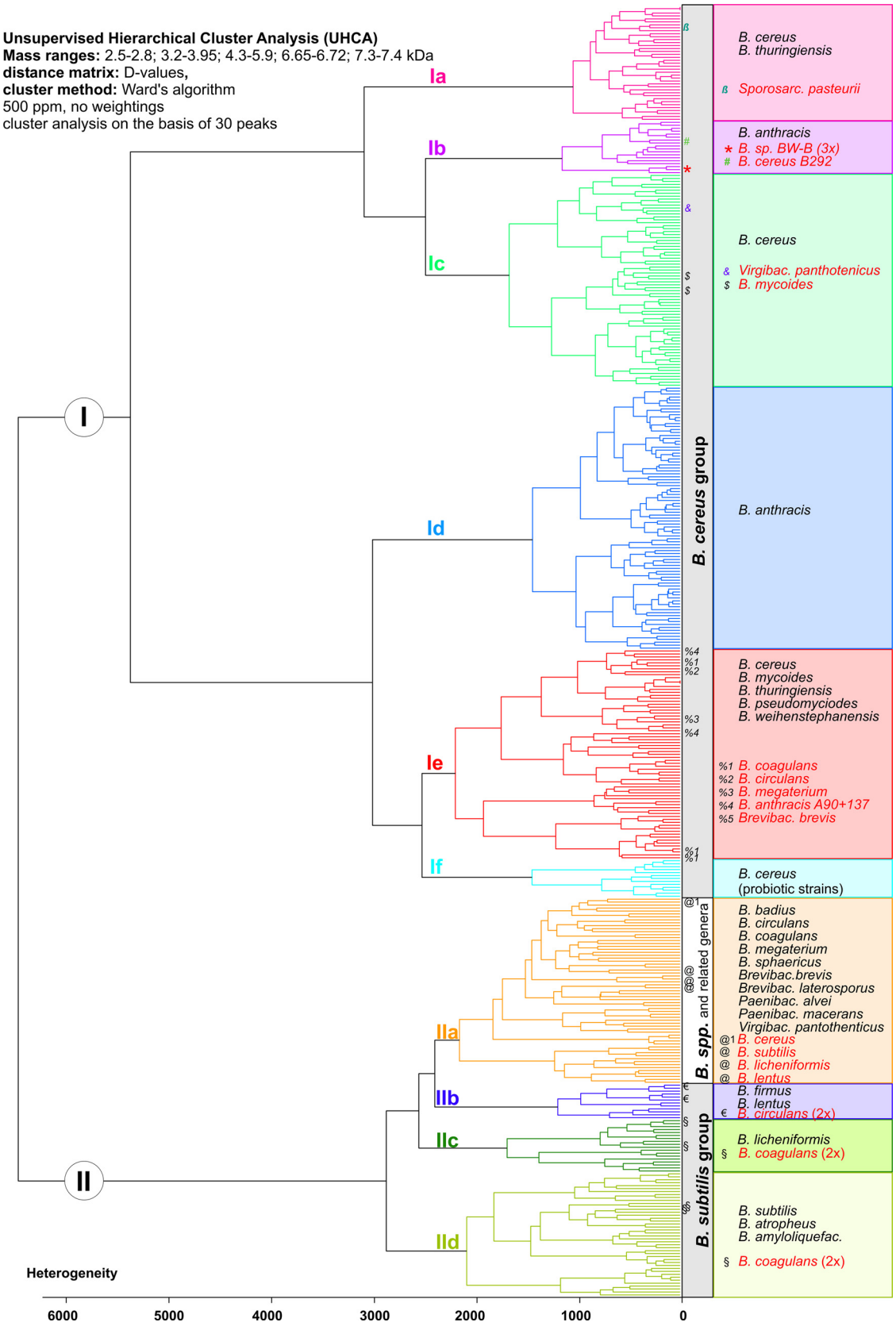


FIG. 5. Hierarchical cluster analysis of microbial mass spectra. Peak tables containing 30 peaks from each mass spectrum were obtained from a total of 437 mass spectra of *Bacillus* species. The peak tables were converted into bar code spectra, which served as inputs for hierarchical clustering on the basis of *D* values and Ward's algorithm. Species names in red denote outliers.

TABLE 2. Confusion matrix showing the classification results of the training data obtained by an optimized ANN[a]

| Class | ANN prediction for class: | | | Sum |
|---|---|---|---|---|
| | i (*B. anthracis*) | ii (*B. cereus* group[b]) | iii (*Bacillus* spp.[c]) | |
| i (*B. anthracis*) | 74 | 0 | 0 | 74 |
| ii (*B. cereus* group[b]) | 0 | 127 | 0 | 127 |
| iii (*Bacillus* spp.[c]) | 0 | 3 | 92 | 95 |

[a] The MALDI-TOF MS-based ANN model was established by training with spectra from three classes: i (*B. anthracis*), ii (*B. cereus* group [except *B. anthracis*]), and iii (*Bacillus* spp. [except *B. cereus* group members]). The classification accuracy in the teaching data for identification of *B. anthracis* was determined to be 100%.

[b] Except *B. anthracis*.

[c] Except *B. cereus* group members.

TABLE 3. Confusion matrix showing ANN classification results of independent test data

| Class | ANN prediction for class: | | | Sum |
|---|---|---|---|---|
| | i (*B. anthracis*) | ii (*B. cereus* group[b]) | iii (*Bacillus* spp.[c]) | |
| i (*B. anthracis*) | 31 | 0 | 0 | 31 |
| ii (*B. cereus* group[b]) | 0 | 52 | 2 | 54 |
| iii (*Bacillus* spp.[c]) | 0 | 3 | 39 | 42 |

[a] The confusion matrix shows the ANN classification results for independent test data. The classification accuracy for class i (*B. anthracis*) was determined to be 100%, as no false-positive or false-negative classifications were found for the class. The values for the identification of class ii (*B. cereus* group [except *B. anthracis*]) and class iii (*Bacillus* spp. [except *B. cereus* group members]) were determined to be 96% (five misclassifications; see the text for details).

[b] Except *B. anthracis*.

[c] Except *B. cereus* group members.

(IId) constitute separate clusters of closely related species with no or few outliers.

**Classification analysis using ANNs.** MLP ANNs were employed in this study as pattern recognition tools for supervised classification analysis. The general strategy of ANN analysis included the procedures of teaching and optimization, in which nonlinear discriminant functions are established, which can be later used to model probabilities of class membership of independent validation data. For this purpose, the spectral database was split randomly into a combined teaching/internal-validation data set and an independent test set for external validation. The first spectral subset contained 296 MALDI-TOF spectra, with 20% of them randomly selected for internal validation. The independent test set consisted of 127 spectra. Furthermore, we defined three categories, or classes, of spectral input patterns, consisting of bar code MALDI spectra of *B. anthracis* (class i), spectra of *B. cereus* group members other than *B. anthracis* (class ii), and mass spectra of non-*B. cereus* group members (class iii).

The classification results achieved by an optimized ANN model for the teaching and the external test data are shown in Tables 2 and 3, respectively. The data suggest that *B. anthracis* is identified with 100% accuracy in the teaching data (see Table 2) and, more importantly, also in the test data (see Table 3), as no false-positive or false-negative classifications were observed. The accuracy values for the identification of members of the other two classes are smaller and vary between 99% in the teaching data and 96% in the test data for classes ii and iii. From the confusion matrix of Table 3, it can also be concluded that the sensitivity of ANN classification of the independent test data was 94.5% for class ii (*B. cereus* group members other than *B. anthracis*) and 92.9% for class iii (non-*B. cereus* group members). The specificities were determined to be 97.2% and 97.6% for classes ii and iii, respectively.

**Discriminating spectral features.** ANN analysis of MALDI-TOF spectral fingerprints is a powerful classification technique, but it has the drawback that the decision rules are not easy to obtain. In order to identify the discriminating spectral information on which the ANN classification relies, we additionally employed univariate statistical methods, such as independent *t* tests. These *t* tests were carried out at any given *m*/*z* value for two-class classification problems, using bar code spectra obtained from peak tables with 30 peaks per spectrum as inputs. The tests were applied as a univariate measure of deviation and provided *P* values, which could be used to assess the discriminative power of the spectral feature under investigation (Fig. 6, curves A to C). Small *P* values cast doubt on the null hypothesis of equal class means.

In order to identify *B. anthracis*-specific biomarkers, univariate *t* tests were applied to a data set in which 105 spectra of *B. anthracis* strains were combined in class 1. Class 2 contained 318 spectra from the 23 remaining species. The *m*/*z* dependence of *P* values in Fig. 6A indicated a number of biomarkers, among them peaks at 3,068, 3,214, 4,606, 5,413, and 6,679 Da. While the potential marker peaks at 5,413 and 6,679 Da could be visually identified from the gel views in Fig. 3 and 4, the identification of the signal at 4,606 Da required a systematic statistical evaluation approach. The comparison of spectra from *B. anthracis* (105 spectra) and *B. cereus* (147 spectra) revealed exactly the same ranking of discriminative mass signals. The *t* tests showed that the peaks at 3,068, 3,214, 4,606, 5,413, and 6,679 Da were highly discriminative for these closely related species. When spectra from the *B. cereus* group (286 spectra; 6 species) were tested against spectra of the remaining species (137 spectra; 18 species, among them many *B. subtilis* group members), the *P* values of the *t* test indicated the presence of *B. cereus* group-specific peaks. The spectra of *B. cereus* group members exhibited mass signals at 3,683, 4,334, 5,171, 5,886, and 7,368 Da, which are typically absent in spectra of species not belonging to the *B. cereus* group of bacilli (Fig. 6C).

**Molecular assignments of biomarkers.** MALDI-TOF MS of bacterial samples offers the possibility to reveal the molecular identities of bacterial protein biomarkers. For example, genome sequences of *B. anthracis* can be utilized to obtain protein sequences and to calculate protein masses. The molecular identities of biomarkers can be then determined by comparing experimental MS data with the molecular masses of proteins in databases such as UniProtKB. Although these types of assignments can only be of a tentative character (and should be carefully interpreted), the possibility of molecular identification of biomarkers constitutes one of the most valuable aspects of the MS-based identification technique.

Table 4 gives an overview of the tentative assignments of 55 typical mass signals in MALDI spectra of *B. anthracis*. The assignments were made by considering posttranslational modifications (excision of N-terminal methionine) and the presence of double-charged protein species.

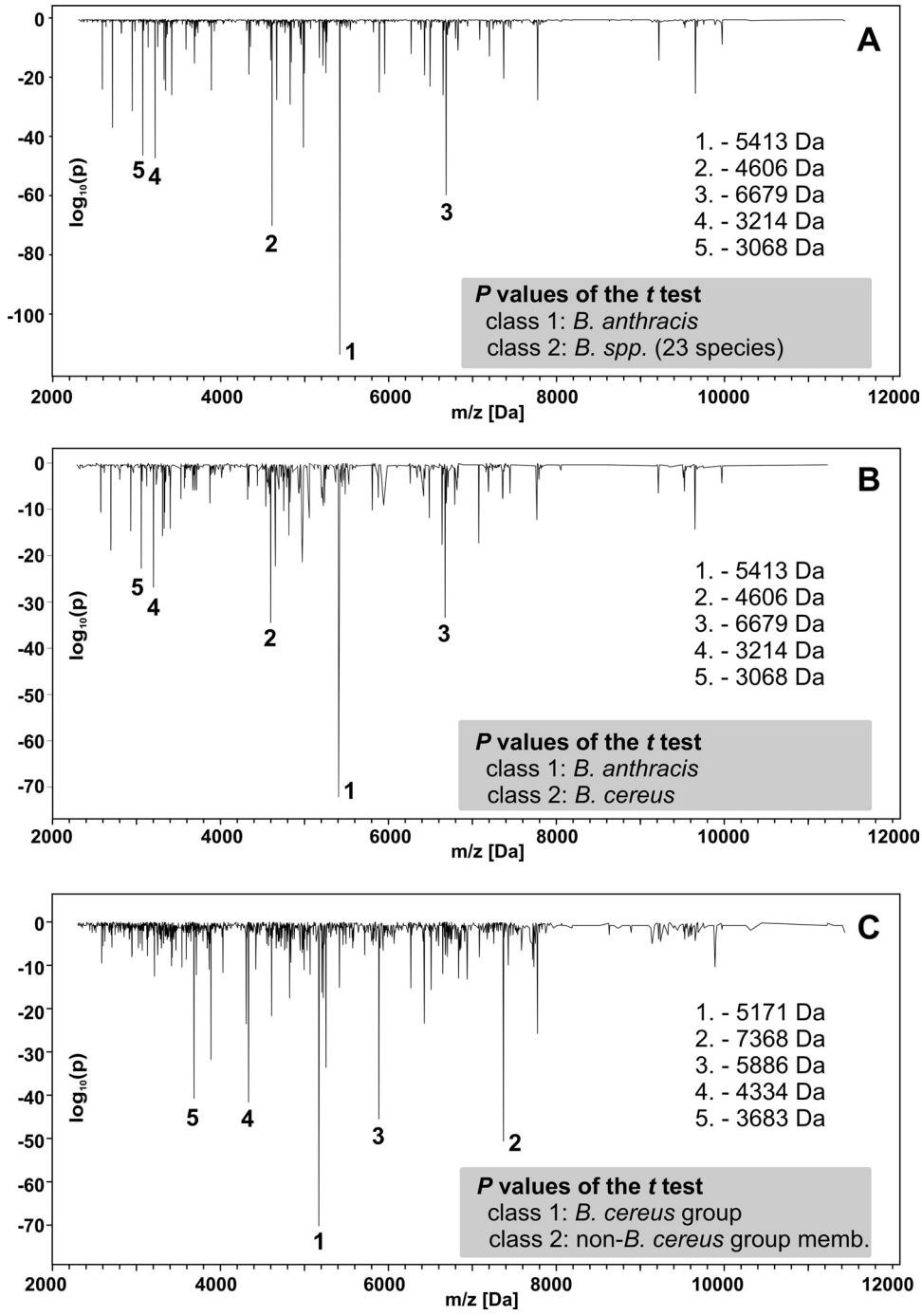


FIG. 6. Identification of discriminating spectral features by independent *t* tests. The *t* tests provided *P* values and were carried out on the basis of peak tables with 30 peaks per mass spectrum. Small *P* values cast doubt on the null hypothesis of equal class means. The smaller the *P* value, the higher the discriminating potential of the respective mass peak. Note the logarithmic scaling. (A) Class 1 contains 105 mass spectra from *B. anthracis*, and class 2 contains 318 spectra from 23 species, including its closest relative, *B. cereus*. (B) *B. anthracis*, with 105 MALDI spectra versus 147 spectra from *B. cereus*. (C) *B. cereus* group (286 spectra) versus non-*B. cereus* group species (137 spectra).

## DISCUSSION

**The importance of sample preparation.** The TFA sample preparation protocol for MALDI-TOF MS-compatible inactivation of highly pathogenic microorganisms permits the reproducible collection of mass spectra from bacterial samples with a large number of mass signals at a low noise level. The two factors—large numbers of mass signals and a high level of reproducibility—are considered equally important for the successful application of the MS technique in microbiological diagnostics. The relatively high number of signals in microbial

TABLE 4. Tentative assignments of *B. anthracis* biomarker peaks

| Observed mass (Da) | Protein in UniProtKB[a] | Description | Predicted MM (Da)[b] |
|---|---|---|---|
| 2,533 | Unassigned | Double-charged species of unassigned peak at 5,066 Da | |
| 2,587 | Q81JG9 | 50S ribosomal protein L34 | 2,586.1[c] |
| 2,630 | Q81Y79 | SASP O | 2,630.3[c] |
| 2,708 | Q81U79 | Putative uncharacterized protein | 2,707.2[c] |
| 2,813 | Unassigned | | |
| 2,944 | Q81LW9 | 50S ribosomal protein L33 2 | 2,943.9[c] |
| 3,068 | Unassigned | | |
| 3,080 | Unassigned | | |
| 3,132 | Q6HUG3 | 50S ribosomal protein L32 | 3,132.2[c] |
| 3,214 | Q81VR2 | 50S ribosomal protein L30 | 3,213.3[c] |
| 3,247 | Unassigned | Double-charged species of unassigned peak at 6,491 Da | |
| 3,322 | Q81LS7 | 30S ribosomal protein S21 | 3,321.9[c] |
| 3,340 | Q81KU1 | SASP B | 3,340.2[c] |
| 3,418 | Q81UL0 | SASP | 3,418.3[c] |
| 3,541 | Q81RQ3 | SASP, α/β family | 3,541.4[c] |
| 3,595 | Q81S96 | Putative uncharacterized protein | 3,595.1 |
| 3,653 | Q6F014 | Putative uncharacterized protein | 3,655.5 |
| 3,683 | Q81YF5 | Cold shock protein CspB | 3,684.0[c] |
| 3,694 | Q81S32 | Putative uncharacterized protein | 3,695.5 |
| 3,885 | Q81V00 | Amino acid permease domain protein | 3,886.7[c] |
| 3,921 | Q6EZR7 | Conserved domain protein | 3,921.7 |
| 4,028 | Unassigned | | |
| 4,334 | Q81VQ6 | 50S ribosomal protein L36 | 4,334.3 |
| 4,451 | Q81ZB4 | Putative uncharacterized protein | 4,452.2 |
| 4,507 | B0AS79 | Conserved-domain protein | 4,511.9 |
| 4,551 | Q81Y87 | SASP N | 4,551.0 |
| 4,606 | Q81VA0 | Putative uncharacterized protein | 4,608.4 |
| 4,664 | Q81KT1 | Putative uncharacterized protein | 4,664.2 |
| 4,797 | Q81QJ2 | Putative uncharacterized protein | 4,798.8 |
| 4,822 | Q6EZN2 | Conserved-domain protein | 4,824.6[c] |
| 4,954 | Q81U02 | Putative uncharacterized protein | 4,954.1 |
| 4,981 | Q81XK6 | Putative uncharacterized protein | 4,982.8[c] |
| 5,066 | Unassigned | | |
| 5,171 | Q81JG9 | 50S ribosomal protein L34 | 5,171.1 |
| 5,215 | Unassigned | | |
| 5,250 | Q81U86 | Putative uncharacterized protein | 5,251.2 |
| 5,259 | Q81Y79 | SASP O | 5,259.7 |
| 5,413 | Q81U79 | Putative uncharacterized protein | 5,413.3 |
| 5,886 | Q81LW9 | 50S ribosomal protein L33 2 | 5,886.8 |
| 5,949 | Q81U56 | Putative uncharacterized protein | 5,951.3 |
| 6,263 | Q6HUG3 | 50S ribosomal protein L32 | 6,263.4 |
| 6,380 | Unassigned | | |
| 6,425 | Q81VR2 | 50S ribosomal protein L30 | 6,425.5 |
| 6,491 | Unassigned | | |
| 6,643 | Q81LS7 | 30S ribosomal protein S21 | 6,642.7 |
| 6,679 | Q81KU1 | SASP B | 6,679.4 |
| 6,835 | Q81UL0 | SASP | 6,835.5 |
| 7,082 | Q81RQ3 | SASP, α/β family | 7,081.7 |
| 7,165 | Q81VR7 | 30S ribosomal protein S14 type Z | 7,165.5 |
| 7,269 | Q81WS1 | Conserved domain protein | 7,267.2 |
| 7,368 | Q81YF5 | Cold shock protein CspB | 7,367.0 |
| 7,771 | Q81V00 | Amino acid permease domain protein | 7,772.4 |
| 9,217 | Unassigned | | |
| 9,649 | Q6EZN2 | Conserved domain protein | 9,648.3 |
| 9,966 | Q81XK6 | Putative uncharacterized protein | 9,964.6 |

[a] The protein identity was determined using a protein database search engine (http://www.uniprot.org/) that allows to access the UniProtKB/SwissProt database.

[b] Molecular mass without N-terminal Met.

[c] Double charged ion.

mass spectra—we typically observe between 50 and 100 per individual spectrum—is in large part due to the ability of TFA to act as an organic solvent, i.e., to effectively solubilize bacterial proteins, preserving the structural integrity (primary structure) of the proteins under investigation (26). In this context, it should be emphasized that the combination of TFA inactivation and MALDI-TOF MS can be considered as effective as ICMS, a simple and widely used mass spectrometric technique in which microbial cells are analyzed directly with only minimal sample pretreatment (14).

A high level of spectral reproducibility, the other crucial aspect of the MS application, was achieved by establishing a rigorous sample preparation protocol that included standardization of cultivation conditions, such as the type of cultivation medium, the cultivation time and temperature, and also the definition of strict parameters for data acquisition.

**Important aspects of spectral preprocessing.** In any spectrometry-based classification analysis, spectral preprocessing aims to improve the accuracy and robustness of subsequent pattern analysis. Also, in MALDI-TOF MS, an optimal preprocessing workflow is the key to reliable classification and included in our study routine tests for spectral quality, normalization, baseline correction, and peak detection. In particular, the strategy of peak detection was found to be highly relevant. Mainly for this reason, the in-house-developed Matlab routines (see Materials and Methods) included specifically designed preprocessing functions, among them a fast and robust peak detection routine. The key feature of the latter routine is a sigmoid function that models the lower analytical sensitivity of the MALDI-TOF MS technique in the high-mass range. Another vital element of our peak detection routine allows the definition of the number of peaks to be determined per mass spectrum. This turned out to be useful for subsequent pattern analysis by UHCA or ANNs. A detailed description of the methods, concepts, and principles of MS preprocessing and fingerprinting will be provided in a separate publication.

In the present work, the general strategy of analysis included statistical evaluation of a large number of microbial mass spectra. Based on a ranking of *P* values obtained in univariate and independent *t* tests, we have been able to determine species (*B. anthracis*)-specific and *B. cereus* group-specific biomarkers suitable for discrimination and identification of the microorganisms of interest (Fig. 6). Here, we discuss the molecular identities of some of these biomarkers.

**Ribosomal biomarkers.** Ribosomal proteins and certain housekeeping proteins have been suggested to be responsible for many mass signals detected by ICMS (12, 27, 34). Since up to 21% of the cell's overall protein content is ribosomal (2) and because of the fact that ribosomal proteins, as part of the cellular translational machinery, are constitutively expressed in vegetative cells, these proteins constitute a stable ensemble of protein biomarkers suitable for use by fingerprinting techniques.

Within the scope of the present work, we found tentative evidence that the *B. cereus* group-specific biomarkers at 4,334, 5,171, and 5,886 Da (Table 4) are due to ribosomal proteins. For example, a search in the UniProtKB/Swiss-Prot database revealed that the 50S ribosomal proteins L36, L34, and L33-2 of *B. anthracis* have molecular weights (MW) of 4,334.3,

5,171.1, and 5,886.8, respectively. Thus, the agreement of sequence information contained in the proteomic database and accurate mass measurements of intact proteins by MALDI-TOF MS allowed us to tentatively assign at least some of the most discriminating *B. cereus* group-specific biomarkers. This preliminary assignment is additionally supported by the results of a BLAST search. For example, the program NCBI BLASTP 2.2.17 revealed for the 50S ribosomal protein L34 of *B. anthracis* (SwissProt entry Q81JG9), with a calculated MW of 5171.1, a total of 25 database entries with 100% amino acid sequence identity. It was interesting that all of these entries originated from strains of *B. cereus* group members, such as *B. anthracis*, *B. cereus*, *B. thuringiensis*, and *B. weihenstephanensis*. Conversely, none of the organisms with less than 100% coverage belonged to the *B. cereus* group. Interestingly, this BLAST search also revealed 86% sequence coverage with the 50S ribosomal protein L34 of *B. licheniformis* DSM 13 (entry Q65CM7). According to the proteomic database, in *B. licheniformis*, the MW of the L34 protein is 5,253, which again fits nicely with our experimental data (Fig. 1D, peak at 5,254 Da).

Although final proof of the peak assignments is not available yet, we found ample evidence suggesting the ribosomal origin of many *B. cereus* group-specific biomarkers. A more detailed analysis could be carried out using tandem MS with some type of fragmentation and peptide analysis in a protein database. With such an approach, it is hoped that the molecular identities of other important biomarkers, such as the *B. anthracis*-specific marker at 5,413 Da (Fig. 2 to 4 and 6A), can also be verified.

**Spore biomarkers.** A closer inspection of the mass spectral profiles of *B. anthracis* and other members of the *B. cereus* group revealed further interesting details. For example, a large number of mass spectra from cultures of *B. cereus* group members exhibited prominent mass peaks at 6,679, 6,695, 6,711, and 6,835 Da. These signals are well known in the literature and have been described as spore biomarkers (6, 11, 17). These spore markers arise from SASPs, a group of proteins present in large amounts in the core regions of *Bacillus* endospores. SASPs play a key role in the protection of the spore's DNA from UV light (32, 33) and, upon germination, serve as a source for amino acids by their degradation (31). Because the amino acid sequences of selected SASPs are species specific, SASP peaks have been suggested as biomarkers for rapid differentiation and identification of spore preparations from *B. anthracis* and *B. cereus* using MS (6, 7, 11, 17, 18).

In agreement with the literature, we found in this work a *B. anthracis*-specific SASP spore marker at 6,679 Da (Fig. 1 to 4). This signal can be predicted from the *sspB* gene of *B. anthracis*, which codes for the α/β-type SASP with a nominal MW of 6,810. As suggested by Castanha et al. and Demirev et al. (6, 9), SASPs may undergo posttranslational modification in which a methionine with a mass of 131 is cleaved, giving the experimental mass of 6,679 Da. Furthermore, mass spectral profiles of *B. cereus* strains displayed two other spore marker peaks, either at 6,695 or at 6,711 Da (Fig. 4). In the present work, the latter peak was also found in *B. mycoides* and *B. thuringiensis*. According to the genome sequences of *B. cereus* (*sasP*-2 gene), these mass signals have been associated with α/β-type SASPs with masses of 6,826 and 6,842 Da (6, 7). Again, cleavage of the N-terminal Met could explain the discrepancy between predicted and experimental masses.

The presence of species-specific SASP peaks correlates well with a number of other mass signals. For example, a species-invariant mass peak at 6,835 Da is found in all spectra of *Bacillus* strains exhibiting one of the SASP peaks at 6,679, 6,695, and 6,711 Da (Fig. 3 and 4). This invariant signal is due to the presence of a second SASP encoded by the *sasP-1* gene with a predicted MW of 6,835 (6,966, Met cleavage) (6). Experimental mass spectra with SASP peaks also show signals of double-charged ions. These signals can be found as species-specific SASP peaks at either 3,340, 3,348, or 3,356 Da and as an invariant SASP at 3,418 Da (Fig. 3).

From our experimental findings and the discussion above, it is quite clear that approximately 50% of the preparations of vegetative cells contained substantial amounts of SASPs. SASPs are known to be indicative of developing or dormant spores, so it was astonishing to us to detect high-intensity SASP peaks in mass spectra of growing cell cultures. On the other hand, the presence of strong SASP signals in mass spectra of *Bacillus* cultures may be the result of the specifically increased sensitivity of the TFA sample preparation technique for these markers. For example, it was reported earlier that acid treatment effectively extracts SASPs from spore preparations (31, 37). Second, SASPs are easy ionizable and are probably preferentially detected by MALDI-TOF MS. Finally, the concentration of SASPs in dormant endospores is substantial and varies between 8 and 20% of the total spore protein content (31). Consequently, it is likely that even small numbers of spores in a large surplus of vegetative cells give intense SASP signals. We have to admit, however, that no data are currently available that could prove these assumptions. Further systematic investigations will reveal at which sporulation stage one can detect SASPs by using MALDI-TOF MS.

Within the context of the present study, which aimed to establish an MS-based classification technique for bacilli, the presence or absence of SASP biomarkers must be considered a confounding factor that is not directly related to taxonomy. On the other hand, as some of the biomarkers have been proven to be species or group specific, SASP features may be beneficial and improve the accuracy of the MS-based detection method.

**UHCA.** UHCA clearly demonstrated that mass spectra of bacilli contain taxonomic information. Although the classification observed with UHCA in the example of 437 individual microbial spectra is not perfect (Fig. 5 shows outliers), the dendrogram allows us to draw a number of important conclusions. First and foremost, UHCA demonstrates the existence of two main spectral groups: in the dendrogram in Fig. 5, the first cluster is formed by *B. cereus* group members and the second contains the remaining species of the genus *Bacillus* plus a few species of related genera, such as *Paenibacillus*, *Brevibacillus*, *Virgibacillus*, and *Sporosarcina*. Second, spectra from *B. anthracis* clearly differed from the spectra of other *B. cereus* group species. For example, the large cluster Id of Fig. 5 contains exclusively mass spectra of *B. anthracis* with no outliers. The second *B. anthracis* cluster, cluster Ib, is formed by mass spectral profiles of *B. anthracis* strains like vollum, A13, A23, and A65. More importantly, this cluster also contains one outlier from *B. cereus* (strain B292) and three replicate spectra of a *B. cereus* group member with unclear species

assignment (strain BWB-B, an environmental isolate). Although we do not have an explanation that could account for the existence of two different *B. anthracis* clusters, it is evident that cluster Ib is composed of spectra with greater similarity to *B. cereus* and other *B. cereus* group members. The situation is further complicated by the presence or absence of spore marker bands (SASPs) that represent an additional source of variance not related to taxonomy in this context.

Analysis of the second cluster, cluster II, also demonstrated the similarity of spectra of *B. subtilis* group members. For example, with a few exceptions, clusters IIb to IId are formed only by species of the *B. subtilis* group. This demonstrates that MALDI-TOF MS, in combination with fingerprinting techniques, can be used to obtain taxonomic information that might be helpful to complement DNA sequencing methods. We are currently investigating the database of *Bacillus* spectra with adapted bioinformatics methods for its applicability to taxonomy. The details of these investigations are beyond the scope of this study and will be provided in a separate publication.

**Supervised classification analysis by ANNs.** The term supervised, or concept-driven, classification analysis is used to describe a group of techniques in which a model is created that maps input objects (spectra) to desired outputs (class assignments) (39). Supervised classification analyses comprise a learning or teaching phase in which a classification function is obtained from class-labeled training data. The performance of the classifier can be subsequently evaluated on the basis of independent test data, which should kept totally separate from the training data. Different types of supervised classification techniques are known, among them the MLP ANNs employed in this study. Within the scope of the present work, we optimized ANN models by training and internal validation. The respective data sets comprised approximately 70% of the bacterial mass spectra. The remaining 30% of the spectra were utilized for objective testing of the ANN model. The results of ANN classification summarized in Tables 2 and 3 demonstrate that MALDI-TOF MS of *Bacillus* can be successfully applied to unambiguously identify strains of *B. anthracis* (class i) or to classify bacilli as members (class ii) or nonmembers (class iii) of the *B. cereus* group. Since the database of MS spectra contains a significant number of spectra with spore biomarkers (SASPs), the ANN classifier can be used to identify MS fingerprints of vegetative cells and mixtures of cells and spores. This is essential, because most of the studies in the field of *B. anthracis* detection by MALDI-TOF MS are limited to the detection of biomarkers from spores.

**Conclusions.** With this comprehensive study carried out on 374 strains of the genus *Bacillus* and a few related genera, among them 102 strains of *B. anthracis*, we have extended previous work in the field of *B. anthracis* detection to vegetative cells. We have been able to detect and to tentatively assign *B. anthracis*-specific and *B. cereus* group-specific protein biomarkers that could be used for objective classification by cluster analysis and ANNs. Our study demonstrates the great potential of MALDI-TOF MS as a rapid, reliable, and objective identification technique for highly pathogenic microorganisms, not only for scientific research purposes, but also under routine conditions. The methodology used in this study can also be applied to identify microorganisms of other genera, such as *Yersinia* or *Burkholderia* (multiplex advantage). Preconditions for successful application of the MS-based technique are (i) rigorous standards for cultivation conditions, (ii) adequate sample preparation (TFA inactivation), (iii) the compilation of databases containing representative numbers of mass spectra, (iv) the application of effective data-preprocessing procedures, and (v) the use of supervised classification models for objective identification.

## ACKNOWLEDGMENTS

We thank R. Reissbrodt (Robert Koch-Institut, Berlin, Germany), B. Niederwöhrmeier (WIS, Munster), and M. Ehling-Schulz (Microbial Ecology Group at the Technical University Munich) for providing *Bacillus* strains. We are grateful to M. Erhard (AnagnosTec, Potsdam), T. Maier and M. Kostrzewa (Bruker Daltonics, Leipzig), and H. Russmann (WIS, Munster) for fruitful discussions and support. Furthermore, the excellent technical assistance of P. Lochau and R. Heinrich (Robert Koch-Institut, Berlin) is acknowledged.

This work was partly supported by the Federal Office of Civil Protection and Disaster Assistance, BBK (BBK F2-440-00-185/04) and WIS (E/E590/4Z015/N5129 and ZVZ-DRMZ-1368-652).

## REFERENCES

1. **Abramova, F. A., L. M. Grinberg, O. V. Yampolskaya, and D. H. Walker.** 1993. Pathology of inhalational anthrax in 42 cases from the Sverdlovsk outbreak of 1979. Proc. Natl. Acad. Sci. USA **90:**2291–2294.
2. **Arnold, R. J., and J. P. Reilly.** 1999. Observation of *Escherichia coli* ribosomal proteins and their posttranslational modifications by mass spectrometry. Anal. Biochem. **269:**105–112.
3. **Bartlett, J. G., T. V. Inglesby, Jr., and L. Borio.** 2002. Management of anthrax. Clin. Infect. Dis. **35:**851–858.
4. **Bavykin, S. G., Y. P. Lysov, V. Zakhariev, J. J. Kelly, J. Jackman, D. A. Stahl, and A. Cherni.** 2004. Use of 16S rRNA, 23S rRNA, and *gyrB* gene sequence analysis to determine phylogenetic relationships of *Bacillus cereus* group microorganisms. J. Clin. Microbiol. **42:**3711–3730.
5. **Beyer, W., C. Bartling, and H. Neubauer.** 2008. *Bacillus anthracis* (Milzbrand), p. 38–55. *In* Mikrobiologisch-infektiologische qualitätsstandards (MiQ) H. 26 Hochpathogene Erreger-Biologische Kampfstoffe, Urban-Fischer, Munich, Germany.
6. **Castanha, E. R., A. Fox, and K. F. Fox.** 2006. Rapid discrimination of *Bacillus anthracis* from other members of the *B. cereus* group by mass and sequence of "intact" small acid soluble proteins (SASPs) using mass spectrometry. J. Microbiol. Methods **67:**230–240.
7. **Castanha, E. R., M. Vestal, S. Hattan, A. Fox, K. F. Fox, and D. Dickinson.** 2007. *Bacillus cereus* strains fall into two clusters (one closely and one more distantly related) to *Bacillus anthracis* according to amino acid substitutions in small acid-soluble proteins as determined by tandem mass spectrometry. Mol. Cell Probes **21:**190–201.
8. **Claydon, M. A., S. N. Davey, V. Edwards-Jones, and D. B. Gordon.** 1996. The rapid identification of intact microorganisms using mass spectrometry. Nat. Biotechnol. **14:**1584–1586.
9. **Demirev, P. A., J. Ramirez, and C. Fenselau.** 2001. Tandem mass spectrometry of intact proteins for characterization of biomarkers from *Bacillus cereus* T spores. Anal. Chem. **73:**5725–5731.
10. **Edwards, K. A., H. A. Clancy, and A. J. Baeumner.** 2006. *Bacillus anthracis*: toxicology, epidemiology and current rapid-detection methods. Anal. Bioanal. Chem. **384:**73–84.
11. **Elhanany, E., R. Barak, M. Fisher, D. Kobiler, and Z. Altboum.** 2001. Detection of specific *Bacillus anthracis* spore biomarkers by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. **15:**2110–2116.
12. **Fagerquist, C. K., A. H. Bates, S. Heath, B. C. King, B. R. Garbus, L. A. Harden, and W. G. Miller.** 2006. Sub-speciating *Campylobacter jejuni* by proteomic analysis of its protein biomarkers and their post-translational modifications. J. Proteome Res. **5:**2527–2538.
13. **Fenselau, C.** 1994. Mass spectrometrry for the characterization of microorganisms. ACS Symposium Series 541. American Chemical Society, Washington, DC.
14. **Fox, A.** 2006. Mass spectrometry for species or strain identification after culture or without culture: Past, present, and future. J. Clin. Microbiol. **44:**2677–2680.
15. **Fritze, D.** 2004. Taxonomy of the genus *Bacillus* and related genera: the aerobic endospore-forming bacteria. Phytopathology **94:**1245–1248.
16. **Granum, P. E., and T. Lund.** 1997. *Bacillus cereus* and its food poisoning toxins. FEMS Microbiol. Lett. **157:**223–228.
17. **Hathout, Y., P. A. Demirev, Y. P. Ho, J. L. Bundy, V. Ryzhov, L. Sapp, J. Stutler, J. Jackman, and C. Fenselau.** 1999. Identification of *Bacillus* spores

by matrix-assisted laser desorption ionization-mass spectrometry. Appl. Environ. Microbiol. **65:**4313–4319.
18. **Hathout, Y., B. Setlow, R. M. Cabrera-Martinez, C. Fenselau, and P. Setlow.** 2003. Small, acid-soluble proteins as biomarkers in mass spectrometry analysis of *Bacillus* spores. Appl. Environ. Microbiol. **69:**1100–1107.
19. **Helgason, E., O. A. Okstad, D. A. Caugant, H. A. Johansen, A. Fouet, M. Mock, I. Hegna, and A. B. Kolstø.** 2000. *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*–one species on the basis of genetic evidence. Appl. Environ. Microbiol. **66:**2627–2630.
20. **Helm, D., H. Labischinski, G. Schallehn, and D. Naumann.** 1991. Classification and identification of bacteria by Fourier-transform infrared spectroscopy, J. Gen. Microbiol. **137:**69–79.
21. **Jackson, P. J., K. K. Hill, M. T. Laker, L. O. Ticknor, and P. Keim.** 1999. Genetic comparison of *Bacillus anthracis* and its close relatives using amplified fragment length polymorphism and polymerase chain reaction analysis. J. Appl. Microbiol. **87:**263–269.
22. **Kennedy, H.** 31 October 2001. Daschle letter bombshell—billions of anthrax spores. New York Daily News.
23. **Keys, C. J., D. J. Dare, H. Sutton, G. Wells, M. Lunt, T. McKenna, M. McDowall, and H. N. Shah.** 2004. Compilation of a MALDI-TOF mass spectral database for the rapid screening and characterisation of bacteria implicated in human infectious diseases. Infect. Genet. Evol. **4:**221–242.
24. **Klee, S. R., M. Özel, B. Appel, C. Boesch, H. Ellerbrok, D. Jacob, G. Holland, F. H. Leendertz, G. Pauli, R. Grunow, and H. Nattermann.** 2006. Characterization of *Bacillus anthracis*-like bacteria isolated from wild great apes from Cote d'Ivoire and Cameroon. J. Bacteriol. **188:**5333–5344.
25. **Krishnamurthy, T., and P. L. Ross.** 1996. Rapid identification of bacteria by direct matrix-assisted laser desorption/ionization mass spectrometric analysis of whole cells. Rapid Commun. Mass Spectrom. **10:**1992–1996.
26. **Lasch, P., H. Nattermann, M. Erhard, M. Stämmler, R. Grunow, N. Bannert, B. Appel, and D. Naumann.** 2008. MALDI-TOF mass spectrometry compatible inactivation method for highly pathogenic microbial cells and spores. Anal. Chem. **80:**2026–2034.
27. **Pineda, F. J., M. D. Antoine, P. A. Demirev, A. B. Feldman, J. Jackman, M. Longenecker, and J. S. Lin.** 2003. Microorganism identification by matrix-assisted laser/desorption ionization mass spectrometry and model-derived ribosomal protein biomarkers. Anal. Chem. **75:**3817–3822.
28. **Priest, F. G., M. Barker, L. W. Baillie, E. C. Holmes, and M. C. Maiden.** 2004. Population structure and evolution of the *Bacillus cereus* group. J. Bacteriol. **186:**7959–7970.
29. **Riedmiller, M., and H. Braun.** 1993. A direct adaptive method for faster backpropagation learning: the RPROP algorithm, p. 586–591. *In* 1993 IEEE International Conference on Neural Networks. IEEE Service Center, Piscataway, NJ.
30. **Schnepf, E., N. Crickmore, J. Van Rie, D. Lereclus, J. Baum, J. Feitelson, D. R. Zeigler, and D. H. Dean.** 1998. *Bacillus thuringiensis* and its pesticidal crystal proteins. Microbiol. Mol. Biol. Rev. **62:**775–806.
31. **Setlow, P.** 1988. Small, acid-soluble spore proteins of *Bacillus* species: structure, synthesis, genetics, function, and degradation. Annu. Rev. Microbiol. **42:**319–338.
32. **Setlow, P.** 2001. Resistance of spores of *Bacillus* species to ultraviolet light. Environ. Mol. Mutagen. **38:**97–104.
33. **Setlow, P.** 2007. I will survive: DNA protection in bacterial spores. Trends Microbiol. **15:**172–180.
34. **Stackebrandt, E., O. Päuker, and M. Erhard.** 2005. Grouping myxococci (*Corallococcus*) strains by matrix-assisted laser desorption ionization time-of-flight (MALDI TOF) mass spectrometry: comparison with gene sequence phylogenies. Curr. Microbiol. **50:**71–77.
35. **Udelhoven, T., M. Novozhilov, and J. Schmitt.** 2003. The NeuroDeveloper: a tool for modular neural classification of spectroscopic data. Chemometrics Intell. Lab. Syst. **66:**219–226.
36. **Ward, J. H.** 1963. Hierarchical grouping to optimise an objective function. J. Amer. Stat Assoc. **58:**236–244.
37. **Whiteaker, J. R., B. Warscheid, P. Pribil, Y. Hathout, and C. Fenselau.** 2004. Complete sequences of small acid-soluble proteins from *Bacillus globigii*. J. Mass Spectrom. **39:**1113–1121.
38. **Wilkins, C. L., and J. O. Lay.** 2005. Identification of microorganisms by mass spectrometry. John Wiley and Sons, Hoboken, NJ.
39. **Wikipedia.** 1 May 2008 posting date. Supervised learning. http://en.wikipedia.org/wiki/Supervised_learning.
40. **World Health Organization.** 2009. Anthrax in humans and animals, 4th ed. WHO, Geneva, Switzerland.